COMPARATIVE RACIAL ANALYSIS OF ENLISTED ADVANCEMENT EXAMS

ITEM DIFFERENTIATION

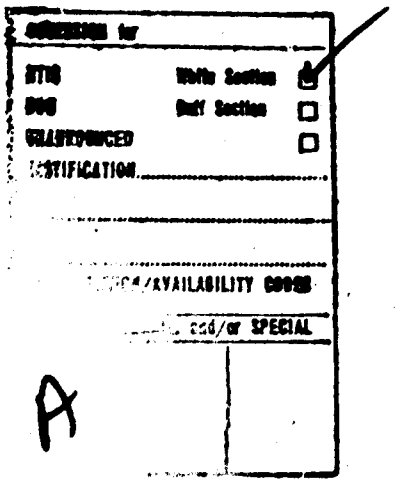
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA

JANUARY 1977

NPRDC TR 77-16

February 1977

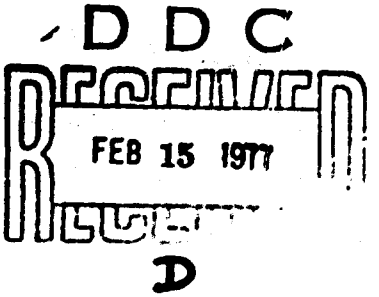
COMPARATIVE RACIAL ANALYSIS OF ENLISTED ADVANCEMENT EXAMS: ITEM DIFFERENTIATION



David W. Robertson
Marjorie H. Royle
David J. Morena

Reviewed by
Martin F. Wiskoff

Approved by
James J. Regan
Technical Director



Navy Personnel Research and Development Center
San Diego, California 92152

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPRDC TR 77-16	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) COMPARATIVE RACIAL ANALYSIS OF ENLISTED ADVANCEMENT EXAMS: ITEM DIFFERENTIATION		5. TYPE OF REPORT & PERIOD COVERED Final Report May 1974 - May 1976 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) David W. Robertson Marjorie H. Royle David J. Morena		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Navy Personnel Research and Development Center San Diego, California 92152		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62763N ZF55.521.031.03.02
11. CONTROLLING OFFICE NAME AND ADDRESS Navy Personnel Research and Development Center San Diego, California 92152		12. REPORT DATE January 1977 13. NUMBER OF PAGES 50
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16 DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Item analysis
Promotion
Racial comparison
Equal opportunity

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

A comparative racial analysis of item differentiation was conducted to determine whether advancement exam items are similarly differentiating between good and poor performers of racial groups. Techniques were also investigated to improve test quality as measured by item differentiation or test reliability. The study specifically investigated (1) the differences in item differentiation between Blacks and Whites, (2) item-difficulty levels that yield maximum item differentiation, (3) the impact on item differentia-

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

tion from constructing tests with particular types of items deleted, and (4) exam construction or processing procedures which would raise test quality for both Blacks and Whites.

Item differentiation levels, calculated as the difference in item-difficulty between high and low scorers (D value) and also as the item-total correlation (r_{it}), were found to be lower for Blacks than for Whites, partly because item-difficulty levels were lower for Blacks. The highest item-differentiation values had corresponding item-difficulty levels which were easier than the median difficulty levels, indicating that the use of easier items should contribute to better item differentiation for both Blacks and Whites. Black-White score differences were reduced by construction of new tests using items of similar difficulty, but test quality was also reduced. Both item differentiation and test reliability were improved by the construction of tests using easier items or more highly correlated items, with slight and varied changes in score differences. The "best" items initially selected by a sequential procedure, applying an internal criterion, were not the same as those selected by an external criterion.

An empirical validation of the present tests on subsequent job performance for both Blacks and Whites was recommended, as was a validation and comparison on internal and external criteria of the alternative test construction procedures identified.

FOREWORD

This study was initiated in response to a request from the Chief of Naval Personnel (Pers-6) to determine the feasibility of developing Enlisted Advancement Exams from items similar in difficulty for both Black and White racial groups, as an approach to improving equal opportunity in career growth for minority groups. Previous studies examined item-difficulty levels both for entire racial groups (Robertson & Royle, 1976--TR 76-6) and for subgroups matched on total test score (Robertson & Montague, 1976--TR 76-34). This report, the third in a series, examines item differentiation and test reliability for the present exams and for modified exams using alternative item selection procedures.

The substantial and valuable assistance of the following persons is gratefully acknowledged: Mr. William E. Montague and DP2 Suzanne Olson, for data processing and computation; and Ms. Hazel F. Schwab, for clerical support.

This study was performed under Exploratory Development Task Area ZF55.521.031 (Career Performance and Selection).

J. J. CLARKIN
Commanding Officer

SUMMARY

Problem

Blacks are advanced to paygrades E-4 and above in smaller proportions than Whites and score lower on the technical knowledge exam than do Whites. It has been found that when exams were constructed only of items similar in difficulty for both Blacks and Whites (to reduce total test score differences), the items were concentrated in the difficult (i.e., guessing) range. This prior finding suggested that such an approach would degrade test quality.

Purpose

As a follow-on, the present study investigated test quality in terms of item differentiation and test reliability. Questions specifically addressed were: (1) what racial differences in item differentiation exist, (2) what levels of item difficulty (P value) yield maximum item differentiation for Black and Whites, (3) what impact constructing tests by selecting particular types of items would have on item differentiation, and (4) what exam construction or processing techniques would raise test quality for Blacks and Whites.

Approach

Item response data for exams of six occupational specialties across four pay grades (i.e., 24 different exams) were analyzed as follows:

1. Racial differences in item differentiation were calculated as (a) the difference in item difficulty between high and low scorers (D value) and (b) the item-total score correlation (r_{it} value).

2. Levels of item difficulty yielding maximum item differentiation were determined by comparing P values with corresponding D and r_{it} values.

3. Three types of modified tests were developed by selecting different types of items: (a) items similar in difficulty for Blacks and Whites (SIM-P), (b) those that were not extremely difficult (UPA-P), and (c) those that were highly correlated (SEQUIN). Black-White score differences in item differentiation and test reliability values for these tests were compared with those for the original test.

4. The SEQUIN item-selection procedure was applied to certain exams using an on-job performance factor as a criterion. Items correlating high with internal (total score) and external (on-job performance) criteria were compared.

Findings

1. Item differentiation was generally lower for Blacks than for Whites, partly because item-difficulty (P value) distributions are lower for Blacks than Whites (p. 7).

2. The highest item-differentiation values (D and r_{it} values) had corresponding item-difficulty levels (P values) that were higher than the median P values (of all items). This indicates that the use of easier items should contribute to higher (i.e., better) item differentiation for both Blacks and Whites (pp. 7 and 11).

3. Selecting items that were similar in difficulty for both Blacks and Whites (SIMP-P test) did reduce mean score differences between Blacks and Whites but it also reduced item differentiation and test reliability. Selecting items that were easier for Blacks (UPA-P test) and those that were highly correlated (SEQUIN test) resulted in slight and varied changes in mean score differences and also increased item differentiation and test reliability (p. 11).

4. The "best" items initially selected by the SEQUIN procedure by applying an internal criterion were not the same as those selected by applying an external criterion. This result raises new questions regarding the relevance of internal-consistency type measures of test quality to measures of subsequent job-relevant performance (p. 14).

Conclusions

1. Item differentiation and test reliability of advancement exams could be improved for both Blacks and Whites by using item selection and construction procedures identified in this study.

2. Developing tests by using items similar in difficulty for Blacks and Whites is not feasible since it reduces test quality. However, developing tests by eliminating excessively difficult items would improve test quality and benefit Blacks.

Recommendations

The empirical validity of the present tests on subsequent job performance should be compared between Blacks and Whites, and the alternative item processing and construction procedures identified herein should be validated and compared on internal and external criteria.

CONTENTS

	Page
INTRODUCTION	1
Problem and Background	1
Purpose	1
METHOD	3
Data	3
Analysis	3
Racial Differences in Item Differentiation	3
Effects of Item-Difficulty (P Value) on Item Differentiation	5
Effects of Item Selection Procedures	5
Effects of Exam Construction and Processing Procedures	6
RESULTS	7
Racial Differences in Item Differentiation	7
Effects of Item-Difficulty (P Value) on Item Differentiation	7
Effects of Item Selection Procedures	11
Effects of Exam Construction and Processing Procedures	14
DISCUSSION	19
Procedures for Improving Advancement Tests	19
Test Validation	19
Identification and Categorization of Valid Items	20
Item Construction Procedures	20
Post Hoc Item Deletion Procedures	21
Balancing Item Biases	22
Implications of the Results	22
CONCLUSIONS	25
RECOMMENDATIONS	27
REFERENCES	29
APPENDIX - METHODOLOGICAL ISSUES IN ITEM ANALYSIS	A 0
DISTRIBUTION LIST	

Page

LIST OF TABLES

1. Advancement Exam Sample Sizes, Means and Standard Deviations by Race 4

2. Range and Median D Values. 8

3. Racial Differences in Item Differentiation for 20 Selected Items of the ADJ3 Exam 9

4. Range and Median of Seven-Item Sets of Highest D Values With Corresponding P Values. 10

5. Mean Total Score, Median P Value, and D Value by Race on Three Types of Tests. 12

6. Reliability, Mean, and Standard Deviation of Four Types of Tests 13

7. Proportions of Theoretical and Applied Type Items in 25 Most and Least Valid Items Selected by SEQUIN (ADJ3 Exam) 15

8. Comparison Between Internal and External Criteria of SEQUIN Item Accretion of Lowest Item-Differentiation Values (ADJ3 Exam). 16

9. Comparison Between Internal and External Criteria of SEQUIN Item Accretion of Lowest Item-Differentation Values (BM2 Exam) 17

FIGURE

1. Illustration of selection of most valid items by SEQUIN (ADJ3 Exam). 14

INTRODUCTION

Problem and Background

The Enlisted Advancement System is one of the Navy's major personnel selection systems being studied to identify and alleviate any condition that might be detrimental to equal opportunity in career growth for all individuals and groups. Advancements to paygrades E-4 and above are competitive and are based on several differentially weighted factors, including the score obtained on a technical knowledge exam, which is substantially weighted. A separate exam, comprising 150 multiple-choice items, is developed for each of approximately 80 Navy ratings (i.e., occupational specialties) and for each paygrade within each rating.

It has been found that Blacks score lower than Whites on the technical knowledge exams, and that a smaller proportion of Blacks than Whites are advanced. To reduce the difference in scores, Robertson and Royle (1975) investigated the feasibility of constructing exams containing only items that were similar in difficulty for both Blacks and Whites. They concluded that the construction of such tests could not be recommended, since the items of similar difficulty were concentrated in the difficult range (i.e., in the guessing range). Although they found that differences in average total test score between Blacks and Whites would be reduced in tests constructed of this type of item, they suggested that such tests would degrade test quality for both groups. Thus, one aspect of the problem is to find ways of constructing advancement tests that provide similar competitive opportunity for all groups, but without loss of test quality, as measured by item differentiation or internal consistency-type reliability.

Purpose

This study investigated racial differences in test quality in terms of item differentiation,[1] including the effects from alternative item selection techniques.

The questions specifically addressed were:

1. What differences in item differentiation exist between Blacks and Whites?

2. What P value levels yield maximum item differentiation for Blacks and Whites?

3. What impact would constructing tests by selecting particular types of items have on item differentiation and test reliability?

4. What exam construction or processing procedures would raise test quality for Blacks and Whites.

---

[1]The term "item differentiation" is used instead of the term typically used in item-analytic studies, "item discrimination," to avoid confusion in the context of racial discrimination.

METHOD

Data

Item response data from the technical knowledge exams of the Series 61 (August 1972) advancement competitions were provided by the Naval Examining Center (now the Naval Education and Training Program Development Center, NETPDC).[2] The ratings selected for analysis were those in which minority group representation was relatively high. The six ratings selected, in competition to paygrades 4 through 7, were:

Aviation Machinist's Mate (Jet Engine Mechanic) (ADJ)
Boatswain's Mate (BM)
Boiler Technician (BT)
Commissaryman (CS)
Hospital Corpsman (HM)
Machinist's Mate (MM)

Data (of Blacks and White only) for the 24 separate competing groups were analyzed. Table 1 presents the sample size, total test mean, and standard deviation for each group.

Analysis

Racial Differences in Item Differentiation

Item differentiation is considered more important than item-difficulty in constructing tests from "good" items; that is, those that are neither extremely easy nor difficult (e.g., P values between 40 and 80) and that relate to the total test score either by a high positive correlation or by higher proportions of high than low scorers answering the item correctly. P values of medium difficulty place upper limits on the relationship of an item to total test score, but do not guarantee effective item differentiation (Nunnally, 1967). The r_{it} and D value statistics were applied to selected items of some of the exams to examine racial differences in item differentiation. The r_{it} statistics were obtained by calculating a Pearson product-moment correlation between each individual's right-wrong response to an item and total test score, yielding a point biserial coefficient. The D value statistic was calculated by rank-ordering total scores and splitting them at the median, creating two subgroups--those who scored high on the total test score and those who scored low. D values were obtained by subtracting the percentage of high scorers who answered the item correctly from the percentage of low scorers who answered the item correctly. Details of these procedures and differences between them are discussed in the Appendix.

---

[2]This data set was also used in previous studies of this series (i.e., Robertson & Royle, 1975 and Robertson & Montague, 1976).

Table 1

Advancement Exam Sample Sizes, Means,
And Standard Deviations by Race

Competition to Pay Grade	Rate	Race					
		Black			White		
		N	$\bar{X}$	SD	N	$\bar{X}$	SD
4	ADJ3	47	52.38	12.60	644	69.96	14.75
	BM3	83	58.07	9.38	1033	64.15	11.86
	BT3	33	61.76	13.37	831	73.77	16.68
	CS3	27	67.59	10.15	447	76.12	11.76
	HM3	104	68.00	11.17	1429	73.45	15.53
	MM3	58	62.48	12.26	1259	72.44	16.56
5	ADJ2	30	58.27	14.39	565	63.55	15.01
	BM2	74	60.12	11.70	569	63.43	10.56
	BT2	28	60.11	10.25	511	73.61	16.57
	CS2	47	64.00	11.41	412	69.01	10.66
	HM2	111	63.60	9.43	1391	70.27	13.40
	MM2	30	56.37	13.69	984	74.09	15.95
6	ADJ1	50	67.78	15.56	400	72.31	15.19
	BM1	115	66.33	11.18	502	72.31	11.49
	BT1	79	70.44	13.57	495	80.70	17.18
	CS1	127	68.27	12.22	661	72.04	11.78
	HM1	26	68.58	6.87	546	71.32	11.08
	MM1	62	62.44	11.26	774	75.39	14.04
7	ADJC	88	66.77	14.23	1014	70.07	14.50
	BMC	193	63.60	12.42	1103	65.75	10.87
	BTC	138	77.91	17.61	956	80.57	15.59
	CSC	165	63.01	14.24	771	65.58	13.92
	HMC	157	71.24	13.73	1817	70.75	13.02
	MMC	110	75.35	13.81	1547	78.73	13.63

Effects of Item-Difficulty (P Value) on Item Differentiation

Although P values of medium difficulty generally produce the most differentiating items, the literature is not in full agreement as to what the ideal P value or range of P values should be. Thus, to investigate the relationship between item-difficulty and item differentiation, D values were rank ordered, seven-item sets were extracted from the top ranks, and the corresponding P values for the D values were identified. Similarily, P values were rank ordered; seven-item sets were extracted from the top, middle, and bottom of the ranks; and the corresponding D values were identified. Finally, r_{it} values were rank ordered, and the P values for the highest and lowest nine r_{it} values were identified. All of the above statistics were computed separately for Blacks and Whites and then compared for racial differences.

Effects of Item Selection Procedures

To compare the impacts on test reliability and item differentiation from alternative methods of item selection, the following three types of tests were simulated and comparative statistics computed:

1. The similar P value (SIM-P) method, developed by Robertson and Royle (1975), which selects only those items having a White P value that is not significantly greater than the Black P value.

2. The upgraded P value (UPA-P) method, developed by Robertson and Royle (1975), which selects only those items having a Black P value greater than 25.

3. The SEQUIN method, developed by Moonan, Balaban, and Geyser (1967), which sequentially identifies and selects items with high correlations to maximize a least squares prediction of a criterion of total score. This "heuristic" method selects items in an "accretion" procedure. The first item selected is the one that correlates most highly with a specified criterion; subsequent items selected are those whose intercorrelations with the items already nominated tend to maximize the correlation coefficient in a regression equation.

Internal consistency reliabilities (Kuder-Richardson type, Ghiselli, 1964, Formula 9-19) were recalculated for the new shortened tests, and compared with those of the original (ORIG) 150-item test. The obtained values for the shortened tests were corrected by the Spearman-Brown Formula (Ghiselli, 1964, Formula 9-4) to provide comparisons of 150-item tests.

Means and standard deviations were recalculated separately for Blacks and Whites for the shortened tests and compared with those of the original test.

Effects of Exam Construction and Processing Procedures

To examine alternative test construction or processing procedures that might raise test quality, a concurrent measure of on-job performance was used. Since no longitudinal type of external criterion was available for the present analysis, such as a measure of technical job performance at the next higher paygrade, the Performance Factor in the composite for advancement competition was utilized for illustrative purposes. (Since this factor is a measure of present rather than subsequent job performance, and includes evaluation of interpersonal behaviors, such as leadership and conduct, in addition to technical effectiveness, its use for illustrative purposes only is emphasized.)

The SEQUIN item-selection procedure was applied to the ADJ3 and BM2 Exams with the Performance Factor as a criterion. Items selected early and late in the sequential procedure by two types of criteria--internal (total score) and external (on-job performance)--were then compared to determine characteristics of valid items in predicting job performance.

RESULTS

Racial Differences in Item Differentiation

Black D values were found to be lower than White D values in 18 of the 24 rate groups (see median difference column of Table 2). A rank order correlation between the median difference and Black sample size of -.42 indicates that the differences are partly attributable to the small Black sample sizes (i.e., the largest differences tend to be associated with the smallest Black samples).

Table 3 illustrates the racial differences in item differentiation in terms of both D value and r_{it} differences for 20 items in the ADJ3 Exam. As shown, Black D values were more than 10 percentage points lower than White D values on 8 items, while White D values were lower on 4 items. (An inspection of all Black-White D value differences revealed that, in 16 exams, Whites were the higher in a majority of those items with differences of at least 10 percentage points; in 2 exams, Blacks were the higher; and in the remaining 6 exams, the frequency with Blacks higher and Whites higher was about equal.) On the ADJ3 Exam, employing the r to Z transformation (Hays, 1963, Formula 15.26.6), Black and White r_{it} values were significantly different for only 12 out of 150 items, which is only 4 items more than would be expected by chance. Of these 12 items, Blacks were lower on 8.

One possible reason for the lower Black item differentiation might be the finding in the Robertson and Royle (1975) study that larger proportions of Black than White P values are concentrated in or near the guessing range (where item differentiation is poorest). The P values for Item 30 in Table 3 tend to support this hypothesis, since the Black P value is in the guessing range, but the P values for Item 16 do not.

Effects of Item-Difficulty (P Value) on Item Differentiation

Since P values of medium difficulty should yield the highest D values, it is of interest to compare the corresponding P values of the highest D values with the median P value of the total test (see Table 4). As shown, the corresponding median P value of the highest D values is higher than the total test median P value in 18 of the 24 rate groups for both Blacks and Whites. (The six exceptions are: Black--CS3, BM2, ADJ1, MM1, BTC, and HMC; and White--MM3, BT2, BT1, HM1, BTC, and HMC.) For example, the corresponding median P value, 42.55, for the highest D values of the ADJ3 Black Group is substantially greater than the total test median P value, 34.0, for that group.

Similar results were obtained from examining the corresponding P values for high and low r_{it} values, and from reversing the orientation and comparing high and low P values and their corresponding D values. These results are presented in greater detail in the Appendix.

Table 2

Range and Median D Values

Rate	Blacks N	Blacks Range	Blacks Median	Whites N	Whites Range	Whites Median	Median Diff.	Rank of Diff.
ADJ3	47	-8.18-64.73	22.83	644	5.46-54.47	24.82	-1.99	15
BM3	83	-9.23-48.25	17.86	1033	3.38-44.55	21.48	-3.62	21
BT3	33	-22.22-69.92	21.11	831	3.84-50.83	25.35	-4.24	23
CS3	27	-26.14-72.73	19.50	447	-0.08-50.05	21.50	-2.00	16
HM3	104	-13.47-48.90	19.51	1429	3.26-44.73	23.46	-3.95	22
MM3	58	-1.43-51.67	20.97	1259	0.35-50.87	24.58	-3.61	20
ADJ3	30	-24.43-75.00	22.62	565	5.89-44.28	24.09	-1.47	13
BM2	74	-11.01-48.96	21.64	569	1.07-39.65	21.11	0.53	5
BT2	28	-23.59-72.31	21.54	511	-6.25-49.55	24.67	-3.13	18
CS2	47	-23.09-63.45	20.05	412	-3.58-37.34	19.05	1.00	3
HM2	111	-17.89-47.89	16.91	1391	-0.03-44.80	22.02	-5.11	24
MM2	30	-19.64-60.00	21.72	984	3.03-50.03	24.86	-3.14	19
ADJ1	50	-17.90-59.03	25.76	400	2.14-52.23	26.06	-0.30	7
BM1	115	-4.48-45.61	22.12	502	2.36-36.97	21.14	0.98	4
BT1	79	-3.23-48.90	23.45	495	0.27-48.84	25.33	-1.88	14
CS1	127	-6.58-48.22	20.86	661	2.32-40.53	21.99	-1.13	9
HM1	26	-28.57-65.00	17.50	546	1.67-44.82	18.84	-1.34	12
MM1	62	-10.71-51.04	22.32	774	-0.75-45.44	24.33	-2.01	17
ADJC	88	-15.80-56.15	24.42	1014	0.79-54.24	25.61	-1.19	10
BMC	193	-2.13-44.97	22.54	1103	-1.21-39.94	20.57	1.97	1
BTC	138	2.18-53.61	23.43	956	1.90-42.13	24.63	-1.20	11
CSC	165	1.33-56.48	22.81	771	-8.92-50.74	22.45	0.36	6
HMC	157	-1.89-51.43	22.05	1817	0.47-46.30	20.66	1.39	2
MMC	110	-22.04-57.91	24.02	1547	0.22-43.14	24.82	-0.80	8

Note. Largest positive difference was assigned Rank 1.

Table 3

Racial Differences in Item Differentiation
For 20 Selected Items of the ADJ3 Exam

Item No.	Black			White			B Minus W Difference	
	P Value	D_s Value	r_{it}	P Value	D_s Value	r_{it}	D_s[a]	Z Test[b]
11	21.28	7.61	.330	34.32	25.42	.287	-17.81	.865
12	31.91	19.93	-.021	26.71	11.10	.036	8.83	-.359
13	42.55	23.73	.346	58.54	20.88	.143	2.85	1.392
14	34.04	41.12	.574	73.45	23.39	.315	17.73	2.101*
15	34.04	7.07	.028	38.51	9.13	.013	-2.06	.096
16	46.81	-1.99	.176	51.55	35.18	.306	-37.17	-.887
17	46.81	32.07	.228	61.49	23.80	.241	8.27	-.089
18	25.53	-1.09	-.026	29.97	8.45	.049	-9.54	-.481
19	27.66	36.21	.126	19.41	17.22	.168	18.99	-.275
20	78.72	34.48	.108	72.98	29.56	.243	4.92	-.896
21	21.28	1.53	.063	23.76	16.68	.074	-15.15	-.071
22	19.15	50.00	.041	36.65	36.58	.343	13.42	-2.031*
23	34.04	25.86	.317	47.36	45.53	.387	-19.67	-.513
24	36.17	40.42	.140	49.07	32.54	.356	7.88	-1.485
25	38.30	9.96	.018	42.86	34.56	.340	-24.60	-2.157*
26	42.55	57.09	.474	49.22	26.60	.272	30.49	1.516
27	29.79	32.76	.387	61.49	26.00	.266	6.76	.871
28	34.04	16.86	.115	44.88	31.50	.182	-14.64	-.440
29	34.04	-1.15	.201	33.54	20.76	.211	-21.91	-.067
30	21.28	10.54	.314	55.75	54.47	.501	-43.93	-1.449

[a]Differences greater than 25.00 are underlined.

[b]Significance of difference between two r_{it} correlations tested using r to Z transformation $\frac{Z_1-Z_2}{\sigma(Z_1-Z_2)}$ (Hays, 1963, formula 15.26.6).

*Two-tail test, $P \leq .05$.

Table 4

Range of Median of Seven-Item Sets of Highest
D Values with Corresponding P Values

Rate	Black				White			
	D Value		P Value		D Value		P Value	
Group	Range	Median	Median of Highest 7 Item	Median of Total Test	Range	Median	Mea. of Highest 7 Items	Median of Total Test
ADJ3	50.00 - 64.73	54.41	42.55	34.04	37.89 - 54.47	41.41	53.42	45.81
BM3	40.36 - 48.25	44.72	45.78	38.74	34.88 - 44.55	35.93	53.44	43.56
BT3	51.47 - 69.92	58.09	45.45	42.42	38.10 - 50.38	41.98	55.48	48.62
CS3	56.67 - 72.73	62.64	44.44	44.44	36.35 - 50.05	39.86	56.38	49.05
HM3	40.52 - 48.90	45.49	48.08	45.19	39.54 - 44.73	42.39	58.43	49.62
MM3	48.33 - 51.67	49.52	44.83	39.66	44.63 - 50.87	46.58	46.62	48.34
ADJ2	53.33 - 75.00	57.47	56.67	36.67	37.81 - 44.28	39.12	49.91	42.52
BM2	41.54 - 48.96	44.15	37.84	39.34	33.36 - 39.65	37.83	46.92	43.24
BT2	50.00 - 72.31	50.26	50.00	39.29	41.63 - 49.55	44.35	47.95	50.20
CS2	49.09 - 63.45	49.82	44.68	42.55	32.36 - 37.34	35.07	54.85	45.63
HM2	37.40 - 47.89	38.62	42.34	41.44	36.08 - 44.80	38.39	50.18	46.30
MM2	49.32 - 60.00	52.78	43.33	36.67	40.51 - 50.03	42.01	58.43	49.54
ADJ1	52.05 - 59.03	53.62	44.00	44.00	43.33 - 52.23	43.73	53.25	48.13
BM1	37.58 - 45.61	41.54	49.57	42.61	34.05 - 36.97	35.23	54.58	45.32
BT1	41.47 - 48.90	44.61	53.16	45.57	43.91 - 48.84	47.09	55.35	55.25
CS1	40.05 - 48.22	41.97	61.42	45.84	36.25 - 40.53	37.62	52.95	50.53
HM1	51.25 - 65.00	53.75	50.00	42.31	35.04 - 44.82	38.74	30.77	45.14
MM1	43.89 - 51.04	45.16	40.32	40.32	39.17 - 45.44	40.70	53.36	48.45
ADJC	47.68 - 56.15	53.14	48.86	44.32	45.85 - 54.24	48.72	51.68	48.72
BMC	38.24 - 44.97	39.09	41.97	41.45	34.24 - 39.94	36.77	51.41	43.70
BTC	47.78 - 53.61	50.86	51.45	51.45	35.78 - 42.13	38.80	50.94	52.41
CSC	44.00 - 56.48	48.00	49.70	43.03	41.08 - 50.74	43.18	50.58	42.93
HMC	46.38 - 51.43	49.03	45.22	46.57	39.35 - 46.30	40.94	46.18	46.46
MMC	43.80 - 57.91	45.13	53.64	50.91	38.47 - 43.14	41.16	54.04	53.04

These results indicate that item differentiation would be improved for both Blacks and Whites by the construction of tests using items that are generally easier, and particularly, with less concentration of items near the guessing range. The results tend to support those of Tinkelman (1971), who proposed a P value of .75 as the optimum average item-difficulty for items with four options, because the error variance due to chance tends to be greater when guessing occurs.

Effects of Item Selection Procedures

Table 5 presents, for five rate groups, the effects on mean score, P value, and D value from employing two types of tests--SIM-P and UPA-P. (The median D value of the SIM-P test is probably an overestimate, and that of the UPA-P test, an underestimate, because each is based on the remaining D values, rather than rescoring section scores and recalculating new D values.) Compared with the original operational tests (ORIG), it was found that:

1. The SIM-P tests substantially reduced Black-White differences in mean score and P value (e.g., for ADJ3 in Table 5, mean score differences were reduced from 17.58 to 3.35; and P value differences, from 11.8 to 3.9) in all five rate groups. However, median D values, as a measure of test quality, were reduced in two of the five Black groups and four of the five White groups (e.g., for HM2, Black median D value remained at 16.9; but that for Whites was reduced from 22.0 to 20.3).

2. The UPA-P tests produced slight and varied Black-White differences in mean score and P value (e.g., for MM3 in Table 5, the mean score difference changed from 9.96 to 9.86), but Black and White median D values all increased (e.g., BM2 Black group, from 39.2 to 46.0).

Table 6 compares the SIM-P, UPA-P, and SEQUIN types of tests with the original tests in regard to test reliability and Black-White mean difference. The SIM-P tests reduced reliability substantially in some rate groups (e.g., for ADJ3, in the corrected r_{xx} column for test length of 150 items, reliability decreased from .863 to .702), and slightly in others (e.g., for BM2, from .729 to .726). The UPA-P and SEQUIN tests both increased reliability slightly. Thus, SIM-P type tests reduced Black-White differences in mean score but at a probably unacceptable cost in reduced test quality for both Blacks and Whites. (The results of the present study, using test quality measures of item differentiation and reliability, provide empirical support for the conclusion of reduced test quality reached in the Robertson and Royle (1975) study.) The effects of UPA-P and SEQUIN tests on Black-White mean score differences are slight and varied. Test quality (i.e., reliability) usually is increased slightly. Such increases in reliability occur most likely because the reliabilities are already quite high--usually in the high .80's. In the one exception, BM2, there is a modest increase from the relatively low .729 to .764 (for UPA-P) and .769 (for SEQUIN).

Table 5

Mean Total Score, Median P Value, and D Value
By Race on Three Types of Tests

Rate Group	Type Test	Black $\bar{X}$ Total	Black Median P Value	Black Median D Value	White $\bar{X}$ Total	White Median P Value	White Median D Value	B Minus W Difference $\bar{X}$ Total	B Minus W Difference Median P Value	B Minus W Difference Median D Value
ADJ3	ORIG[a]	52.38	34.0	22.8	69.96	45.8	24.8	-17.58	-11.8	-2.0
	SIM-P[b,d]	55.16	36.2	20.9	58.51	40.1	21.1	-3.35	-3.9	-0.2
	UPA-P[c,d]	60.19	38.3	24.4	77.52	51.3	24.8	-17.33	-13.0	-0.4
	SIM-P minus ORIG	2.78	2.2	-1.9	-11.45	-5.7	-3.7	---	---	---
	UPA-P minus ORIG	7.81	4.3	1.6	7.56	5.5	0	---	---	---
HM3	ORIG[a]	68.00	45.2	19.5	73.45	49.6	23.5	-5.45	-4.4	-4.0
	SIM-P[b,d]	72.35	48.1	19.5	74.39	50.3	22.7	-2.04	-2.2	-3.2
	UPA-P[c,d]	76.10	49.0	20.3	81.34	53.1	24.6	-5.24	-4.1	-4.3
	SIM-P minus ORIG	4.35	2.9	0	.94	.7	-.8	---	---	---
	UPA-P minus ORIG	8.10	3.8	.8	7.89	3.5	1.1	---	---	---
MM3	ORIG[a]	62.48	39.7	21.0	72.44	48.3	24.6	-9.96	-8.6	-3.6
	SIM-P[b,d]	66.59	41.4	20.1	68.99	44.6	22.7	-2.4	-3.2	-2.6
	UPA-P[c,d]	67.18	41.4	23.3	77.04	50.3	25.8	-9.86	-8.9	-2.5
	SIM-P minus ORIG	4.11	1.7	-.9	-3.45	-3.7	-1.9	---	---	---
	UPA-P minus ORIG	4.70	1.7	2.3	4.6	2.0	1.2	---	---	---
BM2	ORIG[a]	60.12	39.2	21.6	63.43	43.2	21.1	-3.31	-3.9	0.5
	SIM-P[b,d]	61.03	41.2	21.6	61.34	41.5	21.1	-.31	-0.3	0.5
	UPA-P[c,d]	69.27	46.0	24.4	72.24	48.0	22.2	-2.97	-2.0	2.2
	SIM-P minus ORIG	.91	2.0	0	-2.09	-1.7	0	---	---	---
	UPA-P minus ORIG	9.15	6.8	2.8	8.81	4.8	1.1	---	---	---
HM2	ORIG[a]	63.60	41.4	16.9	70.27	46.3	22.0	-6.67	-4.9	-5.1
	SIM-P[b,d]	67.83	45.0	16.9	69.67	45.8	20.3	-1.84	-0.8	-3.4
	UPA-P[c,d]	70.84	45.1	19.0	76.52	49.9	22.3	-5.68	-4.8	-3.3
	SIM-P minus ORIG	4.23	3.6	0	-.6	-.5	-1.7	---	---	---
	UPA-P minus ORIG	7.24	3.7	2.1	6.25	3.6	-.3	---	---	---

[a] Includes the complete set of 150 items.

[b] Includes only items in which the Black P value was not significantly less than than the White P value.

[c] Includes only items in which the Black P value was greater than .25.

[d] Mean total scores are simulated by obtained SIM-P or UPA-P score times $\frac{\text{N items in original test}}{\text{N items in simulated test}}$.

Table 6

Reliability, Mean, and Standard Deviation of Four Types of Tests

Rate Group and N Black/White	Type Test	Reliability			Black		White		B-W
		N[a]	r_{xx}[c]		$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}_\theta$[d]
		Items	Obt.[b]	Cor.[c]					Dif.
ADJ3 47/644	ORIG	150	.863	.863	52.38	12.60	69.96	14.75	-1.285
	SIM-P	74	.538	.702	27.21	5.18	28.86	5.89	-0.298
	UPA-P	114	.830	.865	45.75	11.49	58.92	11.78	-1.131
	SEQUIN	125	.854	.875	44.66	11.21	60.24	13.19	-1.277
	SIM-P minus ORIG			-.161					
	UPA-P minus ORIG			.002					
	SEQUIN minus ORIG			.012					
HM3 104/1429	ORIG	149	.870	.870	68.00	11.17	73.45	15.53	-0.408
	SIM-P	115	.829	.863	55.85	8.47	57.41	11.85	-0.153
	UPA-P	126	.867	.885	64.36	10.18	68.79	14.45	-0.359
	SEQUIN	125	.868	.887	59.69	10.69	64.55	14.48	-0.386
	SIM-P minus ORIG			-.007					
	UPA-P minus ORIG			.015					
	SEQUIN minus ORIG			.017					
HM3 58/1259	ORIG	150	.884	.884	62.48	12.26	72.44	16.56	-0.691
	SIM-P	95	.784	.851	42.17	7.48	43.69	9.78	-0.176
	UPA-P	133	.878	.890	59.57	11.52	68.31	15.48	-0.647
	SEQUIN	125	.879	.897	53.29	10.75	62.29	15.07	-0.697
	SIM-P minus ORIG			.033					
	UPA-P minus ORIG			.006					
	SEQUIN minus ORIG			.013					
BM2 74/569	ORIG	150	.729	.729	60.12	11.70	63.43	10.56	-0.297
	SIM-P	126	.690	.726	51.28	9.98	51.55	9.08	-0.028
	UPA-P	119	.720	.764	54.97	10.55	57.33	9.67	-0.253
	SEQUIN	125	.736	.769	53.23	11.11	56.64	10.02	-0.322
	SIM-P minus ORIG			-.003					
	UPA-P minus ORIG			.035					
	SEQUIN minus ORIG			.040					
HM2 111/1391	ORIG	149	.820	.820	63.60	9.43	70.27	13.40	-0.584
	SIM-P	109	.710	.771	49.66	7.02	51.00	8.92	-0.168
	UPA-P	125	.800	.827	59.43	8.57	64.57	11.84	-0.503
	SEQUIN	125	.814	.840	53.41	8.70	59.52	12.30	-0.581
	SIM-P minus ORIG			-.049					
	UPA-P minus ORIG			.007					
	SEQUIN minus ORIG			.020					

[a] Items remaining after deletion.

[b] Obtained (Obt.) value.

[c] Corrected (Cor.) value for a test of 150 items (Nunnally, 1967, Formula 7-6, p. 223).

[d] X_θ Difference--the mean difference in standard deviation units, calculated by $\frac{\bar{X}_B - \bar{X}_W}{\frac{SD_B + SD_W}{2}}$.

[e] Calculated on Black and White groups combined.

Effects of Exam Construction and Processing Procedures

When the Performance Factor was employed as representative of an external, job-relevant criterion, the SEQUIN procedure reached a maximum validity with a small subset of items. For the ADJ3 Exam, the value of the validity coefficient rose rapidly to a maximum of .206 with the selection of the 20 most valid items (see Figure 1a), then tapered off to a slight negative validity of -.031 for all 150 items. Similarly, for the BM2 Exam, the validity coefficient reached a peak of .273 for 30 items, and a final value of .016. Compared to the validity coefficient, the value of the reliability coefficient, which is largely a function of the number of items in a test, continued to rise steadily (see Figure 1b) during the selection of the first 100 items and leveled off with the selection of the "best" 120 items.

Since SEQUIN also identifies the specific items selected in the "accretion" process, it was possible to categorize items according to content and compare items selected early and late in the process. In the selection of items from the ADJ3 Exam (see Table 7), twice the proportion of theoretical items occurred in the last 25 (i.e., least valid) items as in the first 25 (i.e., most valid), although this 16 percentage point difference was not significant when a chi square test was applied.

Comparing the ADJ3 Exam items selected by both an internal and an external criterion, items with the 14 lowest item-total correlations were identified ($\underline{r}_{it} \leq .050$). With the internal criterion, 11 of the 14 items were among the last third of the items to be selected (see Table 8). However, with the external criterion (the Performance Factor), 12 of the 14 items were in the first third of the items selected. Particularly, three of the items with both a very low $\underline{P}$ value and $\underline{r}_{it}$ value were among those selected earliest--fifth, seventh, and thirteenth--by the external criterion.

Similar results were obtained on the BM2 Exam (see Table 9). Twelve of the 15 items with the lowest item-total correlations were among the first third of items selected by the external criterion, with six of those items among the first 24.

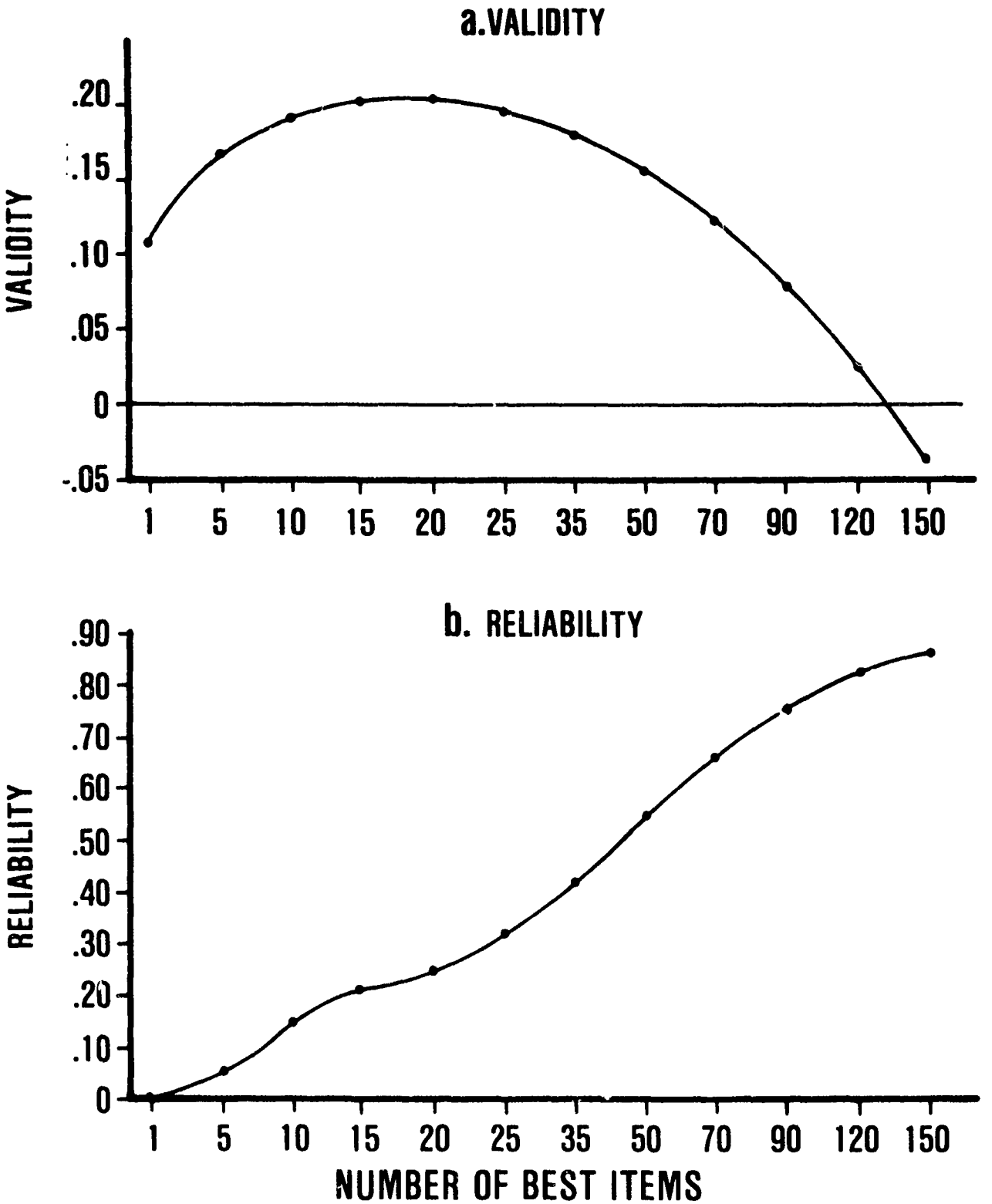


Figure 1. Illustration of selection of most valid items by SEQUIN (ADJ3 Exam.).

Table 7

Proportions of Theoretical and Applied Type Items in 25 Most and Least Valid Items Selected by SEQUIN (ADJ3 Exam)

Item Content Category	Items Selected by SEQUIN: Most Valid 25 Items N	%	Least Valid 25 Items N	%	All Items (150) N	%
Theoretical	4	16	8	32	32	21
Applied	20	80	16	64	110	73
Indeterminant	1	4	1	4	8	5

Note. For a 2 x 2 Matrix of only those items identified as theoretical or applied, $\begin{array}{|c|c|}\hline 4 & 8 \\ \hline 20 & 16 \\ \hline \end{array}$, $\chi^2 = 1.0$, $.50 > p > .25$.

Table 8

Comparison Between Internal and External Criteria Of SEQUIN Item Accretion of Lowest Item-Differentiation Values (ADJ3 Exam)

Item No.	P Value	Lowest r_{it} [a] (≤ .050)	Sequence in which Item was Selected by:	
			Internal Criterion (Total Score)	External Criterion (Performance Factor)
120	.333	.028	61	46
105	.372	.043	65	62
15	.382	.020	92	16
128	.423	.034	103	45
135	.425	-.009	104	17
118	.195	.041	106	39
125	.370	.047	109	4
71	.124	-.022	118	7
55	.534	-.043	123	117
18	.297	.050	128	34
45	.161	-.022	136	13
97	.465	-.054	142	37
12	.271	.023	147	33
113	.100	-.030	150	5

Note. N = 691 (47 Black and 644 White combined).

[a]Values are slight overestimates, since item is included in total score.

Table 9

Comparison Between Internal and External Criteria Of SEQUIN Item Accretion of Lowest Item-Differentiation Values (BM2 Exam)

Item No.	P Value	Lowest r_{it} [a] (≤ .050)	Sequence in which Item was Selected by: Internal Criterion (Total Score)	External Criterion (Performance Factor)
93	.661	.039	52	34
115	.295	.024	59	22
60	.199	.003	99	107
5	.591	.007	100	76
139	.212	-.019	110	43
37	.292	.041	115	14
98	.215	-.060	132	32
107	.104	-.037	137	31
18	.267	-.019	138	136
131	.117	-.090	143	6
81	.152	.031	145	115
19	.093	-.020	147	24
130	.070	.025	148	15
123	.065	-.083	149	62
73	.059	-.048	150	16

Note. N = 643 (74 Black and 569 White combined).

[a]Values are slight overestimates, since item is included in total score.

DISCUSSION

Procedures for Improving Advancement Tests

The problem of how to improve enlisted advancement exams is discussed in the light of the results reported above, the reality of the administration and use of the tests, and the desirability of achieving one or more of three objectives--(1) increasing test reliability, (2) increasing test validity, and (3) decreasing Black-White score differences. It is, of course, easier to state an objective than to achieve it. Even when the rules of good item construction are followed, there is no assurance that the item characteristics desired will be achieved, unless the items are pretested. Nunnally (1967) suggests pretesting at least twice as many items as are intended for the final test. Although such a procedure may be ideal, there are practical limitations in regards to the development of Navy enlisted advancement exams. Advancement is intensely competitive, particularly in the higher paygrades where the proportion of openings is much smaller than the proportion of highly qualified candidates available. If items were pretested on a sample group, the examinees in the sample group might have the advantage of being alerted to the specific content of the forthcoming exam. Also, the P values would probably be lower in the pretest than in the operational test, since the pretest examinees would not be motivated to study as intensely as they would for the operational test.

In lieu of a pretesting procedure, the tests could be improved by the employment of four other procedures:

1. Test validation on an external, job-relevant criterion.

2. Identification of the most and least valid items, and a content categorization of the items identified.

3. Utilization of item construction procedures that tend to produce items with the desired characteristics (e.g., having specified levels of item difficulty, differentiation, and validity).

4. Post hoc item deletion procedures that eliminate undesirable items after administration but prior to final scoring.

Each of these four approaches is discussed in detail below.

Test Validation

The primary concern with a personnel selection test is, of course, its relevance to the purpose of the selection--in the present case, to the individual's effectiveness in the next higher grade for which selected. The measures of test quality investigated in the present study--test reliability and item differentiation--are important to test validity (by setting upper limits on it) but do not of themselves assure test validity.

Validation of the advancement exams on job-relevant criteria is needed for two reasons. First, the courts are becoming increasingly insistent on empirical evidence of the job relevance of personnel selection procedures in compliance with the Civil Rights Act of 1964. Second, CNO Objective Number CNO-1, entitled Retention of Career Personnel (of September 1974), is not addressed to the retention of personnel in general, but rather, to the retention of top quality career personnel. The demonstration of top quality certainly is largely a function of an individual's effectiveness on the job, and motivation to reenlist is certainly heavily influenced by advancement success.

Highly effective validation procedures are available that would be responsive to the above two requirements. The SEQUIN procedure, which was demonstrated with an illustrative job-relevant criterion, was shown to be quite useful, not only to maximize the validity of a test using a subset of items but also to identify the specific items which contribute to, and distract from, prediction of the criterion behavior.

Identification and Categorization of Valid Items

Since SEQUIN identifies the specific item selected in the "accretion" process, it also provides test makers with the capability to analyze and categorize the content of each item. With this knowledge, certain "mixes" of various categories of items could be considered in the construction of future tests. For example, there might be an optimal ratio of theoretical to applied type items for maximum job-relevant validity. The difference between proportions of theoretical and applied items in the first and last 25 items selected in the ADJ3 Exam was not significant. However, with larger pools of items (e.g., the first and last 50-item subsets from a number of exams of similar occupational specialities), significant differences might be identified. Also, categories other than theoretical-applied might be studied, such as the differential validity of the content of the subtest sections.

Item Construction Procedures

In the reliability analysis of five rate groups (see Table 6), the reliability of the BM2 Exam, .729, was substantially below that of the other four groups. This result might be a function of either item statistical or structural characteristics. For example, the median P value (see Table 4) and D value (see Table 2) of the BM2 White group are relatively low among all White groups. (Since the Black and White groups of each rate group were combined to calculate the reliability, the obtained value reflects primarily the distribution statistics of the majority White group.)

Although the literature abounds with guidance for item writing, many of the rules have not been adequately evaluated empirically. In one empirical demonstration of undesirable item characteristics, Dudycha and Carpenter (1973) found that:

1. An inclusive distractor, such as "all (or any or none) of the above" (as opposed to a specific distractor, which is a specified word or phrase) reduces item differentiation.

2. A negative stem structure, which includes the word "not" (as opposed to a positive stem structure, which does not) increases item difficulty.

3. An open-stem structure, which requires the answer to complete the sentence (as opposed to a closed-stem structure, which is a complete sentence) increases item difficulty.

4. The combination of open-positive stems and closed-negative stems in the same test reduces item differentiation.

It was observed that all four of these item designs are used with varying frequency in the present advancement exams, particularly in the BM2 Exam. It would thus be useful to determine whether the use of these (and perhaps other) structures contributes to undesirable item characteristics (e.g., reduced P values or D values).

Also, median P values and D values would probably be increased by raising the criterion values for reuse of items (e.g., P values no less than .30 or greater than .85, and r_{it} with item in score, no less than .05) but subject to item validity with an external criterion.

Post Hoc Item Deletion Procedures

Although pretesting of items is probably not feasible, application of item deletion procedures which eliminate undesirable items (e.g., those with extreme high or low P values, or low differentiation values) subsequent to administration but prior to final scoring for selection purposes might increase the reliability or validity of the exams. The SEQUIN accretion procedures described above demonstrated that a subset of items could be selected that yields a higher validity than, and an equally high reliability as, the total set of items. However, these results should be considered tentative, because the procedure capitalizes on the intercorrelations of the sample data, and is thus influenced by chance. Cross-validation is necessary to ensure that the results are not an effect of sampling error (Henryssen, 1971).

The selection of items to increase reliability will usually tend to increase validity (Henryssen, 1971). However, if excessive emphasis is placed on increasing test homogeneity, the test may become too narrow and one-sided in content to have high validity. In the SEQUIN demonstration with the ADJ3 and BM2 Exams, many of the items with the lowest item-total correlation were selected by an internal criterion near the end of the accretion process, but by an external criterion near the beginning.

A number of reasons might account for these results (other than that the use of the present Performance Factor as an external criterion may not have been appropriate, even for illustrative purposes). If the

test content tends to be heterogeneous, rather than homogenous, as suggested by some of the low intercorrelations among section scores, then internal consistency type measures of reliability may be of limited relevance. This possibility is suggested by a comparison between the reliability and validity coefficients of the ADJ3 and BM2 Exams. Although an internal consistency type measure of reliability places an upper limit on the validity of a test, the situation only applies with homogenous tests. However, with a heterogeneous test, elimination of items with low item-total correlations could result in the reduction of predictable variance. It may be observed that the reliability of the BM2 Exam is lower, but its validity is higher than those of the ADJ3 Exam. Also, when the correlation between two tests is near zero or slightly negative (as is the ADJ3 Exam with the external criterion), the items that correlate lowest with total test score (i.e., the lowest r_{-it} values) could very well be those that correlate highest with an external criterion.

Balancing Item Biases

Another issue pertains to the question of the compatibility of the two objectives identified by the Chief of Naval Personnel to be investigated--the feasibility of compiling "tests composed of questions having identical or correlatable degree of difficulty (Rho) factors for both Blacks and Whites." The Robertson and Royle (1975) study was addressed to the first objective, "identical" difficulty; and the Robertson and Montague (1976) study, to the second, "correlatable" difficulty. The present study addressed both objectives in the context of item differentiation and test reliability.

Both the Robertson and Royle (1975) and the present study found that the construction of tests of items of similar difficulty--from the existing pool of items--was not feasible. The question might be raised as to the existence of, or the possibility of developing, items on which Blacks are superior. If such items were found, tests might be constructed with a "balance" of items in which Whites do well on some, and Blacks, on others. Ironically, such tests would result in increased racial bias, as measured by a decrease in relative item difficulty (Rho value). (The issue of "balancing" item biases is discussed briefly by Cleary and Hilton (1968) and by Jensen (1973).)

Implications of the Results

The demonstrations of improved item differentiation by eliminating excessively difficult items and items with low or negative differentiation suggest the need to implement the item-deletion and item-construction procedures discussed. Such procedures would result in a slight decrease in mean score differences between Blacks and Whites and, in terms of test quality, a slight increase in item differentiation for Whites and a moderate increase for Blacks. Also, any procedure that would raise the level of P values would reasonably be expected to reduce the proportion failed by the exam cut-score, thereby enabling those who passed to continue to compete on their other advancement factors. Although such a procedure was not demonstrated in the present study, it is of particular interest and advantage to Blacks.

However, the SEQUIN demonstration, in which the items selected were compared by internal and external criteria, also suggest that items deleted to increase item differentiation or test homogeneity may be the types of items that best contribute to predicting job-relevant performance by an external criterion. Thus, until external validation studies are performed to determine the relationship of test heterogeneity to subsequent performance in the grade to which advanced, recommendations to implement the procedures discussed above are deemed premature.

CONCLUSIONS

1. Enlisted Advancement Exam item differentiation and internal consistency type test reliability could be improved for both Blacks and Whites by using item selection and construction procedures identified, developed, or demonstrated in this study.

2. The development of tests in which only the items similar in difficulty for both Blacks and Whites are used is not feasible because it would reduce test quality. However, the elimination of excessively difficult items, by either alternative item construction or post-administration item deletion procedures, would improve test quality and, in particular, benefit Blacks, because the proportion of candidates failed by the exam cut-score would be reduced, thereby enabling those who passed to continue to compete on their other advancement factors.

3. The two objectives that were identified for investigation in the present series of studies--the feasibility of compiling "tests composed of questions having identical or correlatable degree of difficulty . . . for both Blacks and Whites"--may not be compatible. As stated above, construction of tests of only items of "identical" difficulty, at least from the existing pool of items, was not feasible. Using "balanced" items might be an alternative to items of "identical" difficulty. However, even if new items could be developed on which Blacks were superior, and tests then constructed with a "balance" of items in which Whites do well on some and Blacks on others, such tests would be characteristic of reduced "correlatable" degree of difficulty. Thus, the use of a measure of relative item difficulty as an indication of possible racial bias appears to be of limited relevance in a study directed towards identifying effective procedures to provide all racial groups with similar opportunities for advancement.

RECOMMENDATIONS

The fundamental question regarding racial differences in advancement should pertain to the relationship of each selection factor, including the present Technical Knowledge Exam, to subsequent job-relevant performance in the grade to which selected. The results of the final phase of the present analysis raise important new questions regarding differences between the "best" items selected by an internal and an external criterion. Thus, implementation of the procedures discussed or demonstrated in the present study (which was at the exploratory level of research), prior to addressing these new questions, would be premature.

It is recommended that: (1) the empirical validity of the present tests on subsequent performance be compared between Blacks and Whites, and (2) the alternative item processing and item construction procedures discussed in the present study be validated and compared on internal and external criteria.

REFERENCES

Cleary, T. A., & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Conrad, H. S. Characteristics and uses of item analysis data. Psychological Monographs, 1948, 62, (Whole No. 295). (See Appendix)

Dudycha, A. L., & Carpenter, J. B. Effects of item format on item discrimination and difficulty. Journal of Applied Psychology, 1973, 58, 116-121.

Ghiselli, E. E. Theory of psychological measurement. New York: McGraw-Hill, 1964.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart, & Winston, 1963.

Henryssen, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Jensen, A. R. An examination of culture bias in the Wonderlic Personnel Test. Paper presented at the National Academy of Sciences, Washington, D.C., October 22, 1973.

Kelley, T. L. Selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 1939, 30, 674-680. (See Appendix)

Lawshe, C. H., Jr. A nomograph for estimating the validity of test items. Journal of Applied Psychology, 1942, 26, 846-849. (See Appendix)

Moonan, W. J., Balaban, J. G., & Geyser, M. J. SEQUIN II: A computerized item selection and regression analysis procedure. San Diego: Navy Personnel Research and Development Center, July 1967. Paper presented at the Military Testing Association Annual Conference, Toronto, Canada, September 1967.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Robertson, D. W., & Montague, W. E. Comparative racial analysis of Enlisted Advancement Exams: Relative item-difficulty between performance-matched groups (Tech. Rep. 76-34). San Diego: Navy Personnel Research and Development Center, March 1976.

Robertson, D. W., & Royle, M. H. Comparative racial analysis of Enlisted Advancement Exams: Item-difficulty (Tech. Rep. 76-6). San Diego: Navy Personnel Research and Development Center, July 1975.

Tinkelman, S. N. Planning the objective test. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

APPENDIX

METHODOLOGICAL ISSUES IN ITEM ANALYSIS

The calculation of item-difficulty and item-differentiation indices for a large number of tests with large subject pools permitted investigation of methodological questions as well as the study of racial group differences.

A number of computational approaches may be used in determining item-differentiation using the item-total relationship, including the r_{it} and D value[1] techniques employed in this study. These and other alternative procedures provide much the same information. The rankings of item-differentiation values by alternative procedures usually yield correlations among the ranks in the .90's (Nunnally, 1967). In computing item-differentiation statistics, if the item itself is included in the total (or section) score, some portion of the correlation value obtained will be an artifact from the presence of the item itself (Nunnally, 1967). (Obviously, the size of this artifact will vary inversely with the number of items in the test/section.) Also, if a test contains subtests (i.e., "sections") of differing content (i.e., a nonhomogenous type test), it may be more appropriate to compare item responses with the subtest score than with total score.

Alternative Item Analysis Procedures Employed

To investigate the effects of including the item in the total score and of computing item-differentiation statistics on sections vice total test scores, the following alternative statistics were computed:

1. r_{is} (w/ item)--item-section correlation, with the item included in the section score.

2. r_{is} (w/o item)--item-section correlation, without the item included in the section score.

3. r_{it} (w/ item)--item-total correlation, with the item included in the total score.

4. r_{it} (w/o item)--item-total correlation, without the item included in the total test score.

---

[1]The D value of the present study is to be distinguished from the Lawshe (1942) D value, adopted from the Kelley (1939) technique, which expresses the difference between the two scoring groups in terms of sigma units.

5. $\underline{D}_s$ value (w/ item)--percentage difference between high and low section scorers who answered the item correctly.

6. $\underline{D}_t$ value (w/ item)--percentage difference between high and low total scorers who answered the item correctly.

$\underline{D}_s$ values (hereafter referred to as $\underline{D}$ values) were calculated on all items for all 24 rate groups employing the above procedure 5. Although this procedure produces values that are overestimates from the presence of the item itself in the section score, it was considered useful for the present analysis, since the primary interest concerned the relative size of the values between Blacks and Whites, rather than the absolute size of the $\underline{D}$ value.

Intercorrelations among section and total test scores were calculated for four selected rate groups.

Effects on Item-Test Correlation From Including Item in Score

Table A-1 presents item-score point biserial correlations for all four alternative responses for seven selected items of the HM3 Exam, calculated both with and without the item included in the score. The correlations between each alternative item response and test score were found to be higher when the item was included in the score than when it was not included. This finding is consistent with discussions in the general literature (e.g., Nunnally, 1967). Inclusion of the item in the section score frequently increases substantially the $\underline{r}_{is}$ of the correct response alternative (e.g., for Item 2 alternative 3, from .211 to .424 for Blacks, and from .095 to .379 for Whites). Inclusion of the item in total score, however, usually increases the $\underline{r}_{it}$ by only .02 to .04 correlation points (e.g., for Item 130 alternative 1, from .235 to .275 for Blacks, and from .188 to .219 for Whites). The increase in $\underline{r}_{is}$, from inclusion of the item in the section score, is greatest in the lowest $\underline{r}_{is}$ values without the item (e.g., for Whites, from .055 to .215 in Item 150, compared with .391 to .449 in Item 30), although the difference in $\underline{r}_{is}^2$ is slight (e.g., for Whites, from .003 to .046, a difference of .043 in Item 150, compared with a difference of .049 in Item 30).

In calculating $\underline{D}$ values, a similar procedure could have been applied by dividing the group into high and low scorers for each item on the basis of their score without that item included. This lengthy procedure was not applied, therefore all obtained $\underline{D}$ values can be considered to be overestimates.

Table A-1

Comparison of Four Methods of Calculating Item-Score Correlations
On Seven Selected Items of the HM3 Exam

Item No.	Item in Score[a]	Response Alternative	Black								White							
			Section (r_{is})				Total (r_{it})				Section (r_{is})				Total (r_{it})			
			1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
2	w/		009	-431	424	130	-019	-155	160	057	-060	-308	379	-046	-029	-101	111	024
	w/o		039	-280	211	168	-014	-129	125	061	-014	-083	095	001	-024	-077	082	028
20	w/		310	-044	-274	-130	289	030	-268	-173	486	-202	-406	-037	437	-174	-370	-032
	w/o		216	-009	-219	-100	247	046	-244	-161	427	-181	-368	-011	411	-165	-354	-021
30	w/		-219	307	-153	-147	-114	212	-207	-014	-151	449	-198	-322	-150	417	-181	-297
	w/o		-187	216	-107	-095	098	170	-187	010	-132	391	-164	-289	-142	392	-166	-282
80	w/		028	-128	257	-237	-036	-177	314	-257	-086	-070	332	-264	-081	-077	271	-210
	w/o		051	-121	116	-109	-030	-176	275	-221	-071	-064	233	-177	-077	-075	243	-186
90	w/		-017	-119	120	-024	-036	-111	136	-038	026	-060	253	-206	039	-074	221	-170
	w/o		004	-076	-024	075	-030	-098	094	-009	042	-021	147	-140	043	-064	192	-151
130	w/		376	-103	-168	-173	275	-199	-002	-108	313	-132	-155	-110	219	-085	-127	-068
	w/o		248	-040	-120	-139	235	-181	014	-097	198	-082	-113	-058	188	-072	-115	-054
150	w		265	-031	-200	-031	235	-014	-071	-162	215	-128	-011	-021	191	-145	-021	033
	w/o		118	053	-190	-018	209	001	-068	-161	055	-036	002	-002	172	-134	-019	035

Note. Correct response is underlined. Decimal points of point biserial correlations have been omitted.

[a]Total or Section score calculated:

w/ --with item in the score
w/o--without item in the score

Comparison of Item Differentiation by Section and Total Score

As expected, $\underline{D}_s$ values were found to be higher than $\underline{D}_t$ values. As illustrated with 15 selected BM2 items in Table A-2, Black $\underline{D}_s$ values exceeded $\underline{D}_t$ values by 4 to 41 percentage points with four exceptions (e.g., in Item 10, the $\underline{D}_s$ value was lower by about 10 points). Also the rank order of item differentiation varied considerably both by method ($\underline{D}_s$ and $\underline{D}_t$) and by race.

Table A-3 presents the item-score correlations, of the correct response only, for 13 items (including the 7 items in Table A-1) from the HM3 Exam, along with corresponding $\underline{D}$ values and $\underline{P}$ values.[2] The ranks (among the 13 items) of alternative item-differentiation values are quite similar across method (e.g., $\underline{r}_{is}$ and $\underline{D}_s$, $\underline{r}_{is}$ and $\underline{r}_{it}$, etc.) when both methods include the item in the score, and when both methods exclude the item. However, the ranks vary when one method with the item included is compared with another method with the item excluded. For example, on Item 110, the White group ranks for $\underline{r}_{is}$ (rank 11) and $\underline{D}_s$ (rank 12), with the item in the score, are nearly the same compared to the $\underline{r}_{is}$ rank without the item (rank 6).

Of particular interest in Table A-3 is the comparison between $\underline{r}_{is}$ and $\underline{r}_{it}$ values (without the item included in the score). If the total test contains section of differing content, use of $\underline{r}_{is}$ may be more appropriate than $\underline{r}_{it}$ (as discussed on page A-1). Tables A-4 and A-5 present intercorrelations among section and total scores for two exams. For example, on the HM3 Exam (see Table A-4), section-section correlations range from -.011 (sections 1 and 6) to .431 for Blacks, and from .019 to .648 for Whites. Section-total correlations range from .363 to .814 for Blacks, and from .370 to .904 for Whites. (The section-total correlations are spuriously high, since the section is included in the total score.)

---

[2]The measure of item-difficulty employed in this item-analysis was the $\underline{P}$ value, the percentage of a group which answers the item correctly (i.e., as defined by Tinkelman (1971, p. 62), the lower the $\underline{P}$ value, the more difficult the item). This measure is to be distinguished from an alternative measure of item-difficulty, Delta value, designated by the Greek letter "Δ," and characterized by higher Δ values associated with more difficult items. This latter measure employs "transformed criterion-scores" of the persons attempting the item and is particularly appropriate in tests measuring speed of performance (Conrad, 1948). Because both Blacks and Whites tend to complete the entire test, the simpler $\underline{P}$ value was used in the present analysis.

Table A-2

Comparison of Two Methods of Calculating Item Differentiation of 15 Selected Items of the BM2 Exam

Item No.	On Section Score				On Total Score			
	Black		White		Black		White	
	$\underline{D}_s$ Value	Rank	$\underline{D}_s$ Value	Rank	$\underline{D}_t$ Value	Rank	$\underline{D}_t$ Value	Rank
10	11.23	114	26.71	27	21.54	26	12.77	64
20	9.53	119	22.70	58	15.60	52	14.96	45
30	32.46	32	16.62	102	13.55	62	5.96	111
40	7.89	125	23.50	53	11.28	75	19.18	19
50	33.55	29	33.55	6	6.45	102	24.13	5
60	9.67	118	9.21	135	-1.17	135	-1.78	143
70	11.79	109	13.26	114	17.22	44	15.97	38
80	11.31	113	17.40	93	4.76	112	2.58	131
90	9.23	120	10.40	129	9.38	84	7.46	97
100	15.31	96	15.31	105	9.89	82	3.51	126
110	38.14	13	15.27	106	24.47	19	14.01	52
120	12.43	105	21.57	71	7.62	97	11.09	74
130	11.40	111	5.14	146	4.25	116	2.35	134
140	40.77	9	27.68	20	-1.83	137	10.78	76
150	28.72	44	8.48	137	10.77	79	2.49	133

Note. Highest $\underline{D}$-value was assigned Rank 1.

Table A-3

Comparison of Four Item Statistics on Selected Items of the HM3 Exam

Item No. (and Test Section No.)	Item in Score[a]	Black				White			
		r_{is}[b]	r_{it}[b]	D_s[b]	P	r_{is}[b]	r_{it}[b]	D_s[b]	P
1	w/	393 [3]	158 [8]	28.34 [4]	28.9	388 [3]	104 [9]	33.38 [3]	48.4
(1)	w/o	139 [6]	118 [8]			081 [8]	072 [9]		
2	w/	424 [2]	160 [7]	30.58 [2]	19.2	379 [4]	111 [8]	30.73 [4]	31.4
(1)	w/o	211 [5]	125 [7]			095 [7]	082 [8]		
3	w/	500 [1]	119 [10]	42.95 [1]	41.4	299 [7]	032 [13]	24.96 [5]	45.0
(1)	w/o	239 [2]	075 [10]			014 [13]	000 [13]		
20	w/	310 [5]	289 [2]	15.71 [7]	49.0	486 [1]	437 [1]	43.44 [1]	59.2
(2)	w/o	216 [3.5]	247 [2]			427 [1]	411 [1]		
30	w/	307 [6]	212 [6]	14.86 [10]	63.5	449 [2]	417 [2]	33.57 [2]	69.7
(2)	w/o	216 [3.5]	170 [6]			391 [2]	392 [2]		
60	w/	294 [7]	073 [12]	19.55 [5]	56.7	239 [10]	041 [12]	19.40 [9]	52.9
(3)	w/o	084 [10]	029 [12]			040 [11]	008 [12]		
70	w/	183 [12]	043 [13]	3.33 [13]	26.9	223 [12]	089 [11]	17.28 [11]	32.2
(3)	w/o	009 [11]	003 [13]			037 [12]	059 [11]		
80	w/	257 [9]	314 [1]	9.91 [11]	35.6	332 [5]	271 [3]	23.41 [7]	30.9
(4)	w/o	116 [8]	275 [1]			233 [3]	243 [3]		
90	w/	120 [13]	136 [9]	15.58 [8]	34.6	253 [9]	221 [7]	17.89 [10]	35.7
(4)	w/o	-024 [13]	094 [9]			147 [5]	192 [4]		
110	w/	239 [10]	098 [11]	19.44 [6]	34.6	229 [11]	126 [7]	16.70 [12]	29.3
(5)	w/o	101 [9]	056 [11]			120 [6]	097 [7]		
130	w/	376 [4]	275 [3]	28.98 [3]	33.7	313 [6]	219 [5]	24.01 [6]	42.4
(5)	w/o	248 [1]	235 [3]			198 [4]	188 [5]		
140	w/	210 [11]	239 [4]	8.69 [12]	24.0	291 [8]	092 [10]	23.10 [8]	24.3
(6)	w/o	-009 [12]	202 [5]			069 [9]	064 [10]		
150	w/	265 [8]	235 [5]	15.07 [9]	9.6	215 [13]	191 [6]	10.48 [13]	10.6
(6)	w/o	118 [7]	209 [4]			055 [10]	172 [6]		

Note. Decimal points of r_{is} and r_{it} point biserial correlations have been omitted.

[a]Total or section score calculated:

w/ --with item in the score
w/o--without item in the score

[b]The rank (among the 13 items only) of each value is indicated by the smaller numbers, which are in superscript, highest value with rank 1 (e.g., for Item 20, White r_{is} of .427, calculated without the item in section score, is rank 1.

Table A-4

Distribution Statistics and Intercorrelations Among Section and Total Scores of the HM3 Exam

			Black				
Section	1	2	3	4	5	6	Total
1		308	001	122	183	-011	383
2			283	364	431	114	814
3				163	152	158	456
4					419	248	684
5						101	701
6							363
Mean	4.61	21.33	9.98	15.29	10.64	5.71	68.00
S.D.	1.73	4.54	2.35	3.32	3.41	1.96	11.17

			White				
Section	1	2	3	4	5	6	Total
1		273	158	192	219	109	370
2			433	629	648	255	904
3				358	352	142	573
4					542	213	797
5						233	800
6							388
Mean	5.15	21.22	10.58	16.63	11.63	6.19	73.45
S.D	1.60	5.29	2.50	4.37	4.07	1.92	15.53

Note. Decimal points for correlations have been omitted.

Table A-5

Distribution Statistics and Intercorrelations Among Section and Total Scores of the BM2 Exam

Black

Section	1	2	3	4	5	6	7	8	Total
1		236	191	363	349	200	057	374	550
2			432	241	295	190	357	404	724
3				253	294	438	320	159	721
4					322	234	259	284	582
5						125	055	274	544
6							305	191	543
7								017	503
8									527
Mean	7.47	13.58	12.35	5.39	5.64	5.32	5.39	4.97	60.12
S. D.	2.31	3.45	3.59	2.09	1.96	1.95	2.09	1.79	11.70

White

Section	1	2	3	4	5	6	7	8	Total
1		257	239	204	174	157	242	217	564
2			354	224	150	221	289	144	650
3				240	239	201	273	255	694
4					152	082	236	163	515
5						093	146	201	452
6							172	137	421
7								256	580
8									500
Mean	8.26	13.71	12.63	5.78	5.78	5.83	6.07	5.31	63.43
S.D.	2.39	3.02	3.25	2.23	1.93	1.75	2.18	1.95	10.56

Note. Decimal points for correlations have been omitted.

It might be reasonable to assume that, if the section-total correlation is low, $\underline{r}_{is}$ would be higher than and more appropriate than $\underline{r}_{it}$ (if the section content is assumed to be homogenous). However, these assumptions are not supported by the few illustrative items of the HM3 Exam in Table A-3. For example, for Blacks, $\underline{r}_{it}$ is higher than $\underline{r}_{is}$ on the two items (140 and 150) from section 6, although this section had the lowest section-total correlation (.363 in Table A-4). Of the two items (20 and 30 in Table A-3) from the section with the highest section-total correlation (.814 in Table A-4), one $\underline{r}_{it}$ is higher, and the other is lower than $\underline{r}_{is}$.

In the light of varying differences between $\underline{r}_{is}$ and $\underline{r}_{it}$, and among section-total correlation (including, quite likely, even sections of heterogeneous content), generally, the most useful measure of item differentiation appears to be $\underline{r}_{it}$ (without the item included in total score). (Nonetheless, use of $\underline{D}_s$ with item in section score is considered useful and adequate for analyzing the relative differences between racial groups in the present study.)

Relationship Between P and D Values

When the corresponding P values for the highest D values were examined (see Table 4 and page 7), the median P value of the highest D values was generally higher than the total median P value. Similar results were also obtained with the corresponding P values for the highest $\underline{r}_{it}$ values in Table A-6. With one exception (the MM3 Black group), these corresponding P values are higher than the total test median P value. For example, the corresponding median P value, 54.19, for the highest $\underline{r}_{it}$ values of the ADJ3 White group, is substantially higher than the total test median P values, 45.81 for that group.

Table A-3 also provides examples of high P values which yield high or low differentiation values (e.g., for the White group, Item 20 P value of 59.2 with $\underline{r}_{it}$ without the item in score of .411, but Item 60 P value of 52.9 with $\underline{r}_{it}$ of only .008), and low P values which yield high or low differentiation values (e.g., Item 80 P value of 30.9 with $\underline{r}_{it}$ of .243, but Item 70 P value of 32.2 with $\underline{r}_{it}$ of only .059.

Reversing the orientation and comparing P values with corresponding D values yielded similar results (see Table A-7). The P values of middle difficulty (e.g., ADJ3 Black group, median P value of 34.04) yield corresponding D values (e.g., 41.12) which are substantially lower than the highest D values (e.g., ADJ3 Black median, 54.41, in Table 4, page 10). Figures A-1, A-2, and A-3 display the median P values and corresponding D values for the 7-item ranked sets of items in Table A-7. It may be observed that the highest P values yield corresponding D values which are higher than the corresponding D values of the lowest P values for both Blacks and Whites.

Table A-6

Range and Median of Nine-Item Sets of Highest and Lowest $\underline{r}_{it}$ Values
And Corresponding $\underline{P}$ Values for Four Rate Groups

Rate	Highest (H) and Lowest (L) $\underline{r}_{it}$	Black				White			
		Ranked $\underline{r}_{it}$		Corresponding $\underline{P}$		Ranked $\underline{r}_{it}$		Corresponding $\underline{P}$	
		Range	Median	Range	Median	Range	Median	Range	Median
ADJ3	H	.440 - .574	.474	21.28 - 51.06	38.30	.376 - .501	.384	29.04 - 81.83	54.19
	L	-.185 - -.108	-.166	17.02 - 57.45	36.17	-.034 - .021	-.019	10.09 - 53.11	38.51
HM3	H	.347 - .511	.371	24.04 - 69.23	47.12	.436 - .481	.446	41.15 - 72.92	58.43
	L	-.253 - -.032	-.117	7.69 - 59.62	29.33	-.092 - .011	-.015	7.42 - 77.75	32.26
MM3	H	.418 - .583	.458	22.41 - 51.72	36.21	.438 - .493	.467	44.00 - 66.16	50.28
	L	-.199 - -.071	-.147	15.52 - 70.69	36.21	-.083 - .041	.007	9.13 - 49.09	30.18
BM2	H	.364 - .559	.398	17.57 - 70.27	47.30	.266 - .370	.287	35.68 - 62.74	49.03
	L	-.150 - -.027	-.079	9.46 - 68.92	20.72	-.088 - -.021	-.053	5.10 - 35.33	11.95

Table A-7

Range and Median of Seven-Item Ranked Sets of Highest, Middle And Lowest P Values and Their Corresponding D Values

Rate Group		Black						White					
		Highest		Middle		Lowest		Highest		Middle		Lowest	
		Range	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range	Median
ADJ3	P	57.45-78.72	59.57	34.04-34.04	34.04	8.51-14.89	12.77	75.31-83.23	79.81	45.34-47.36	46.74	10.09-18.32	14.44
	D	20.73-45.82	27.36	-1.15-44.02	41.12	-7.51-9.88	4.79	20.70-25.09	22.74	14.53-45.53	26.24	5.46-16.45	13.18
HM3	P	76.92-95.19	81.73	44.23-46.15	45.19	5.77-9.62	7.69	80.97-96.36	86.00	48.92-50.80	49.69	7.42-12.60	8.75
	D	1.90-26.58	15.30	14.88-46.40	20.56	-2.42-15.02	9.29	4.93-21.67	12.15	17.58-36.88	30.41	5.90-12.88	9.26
MM3	P	70.69-82.76	74.14	39.66-39.66	39.66	6.90-15.52	15.52	72.99-88.64	79.03	47.74-49.17	48.77	9.13-15.89	13.34
	D	5.71-29.52	18.99	13.10-28.73	14.52	-.24-20.97	6.19	4.44-27.06	19.30	20.80-39.93	28.10	.35 14.88	10.12

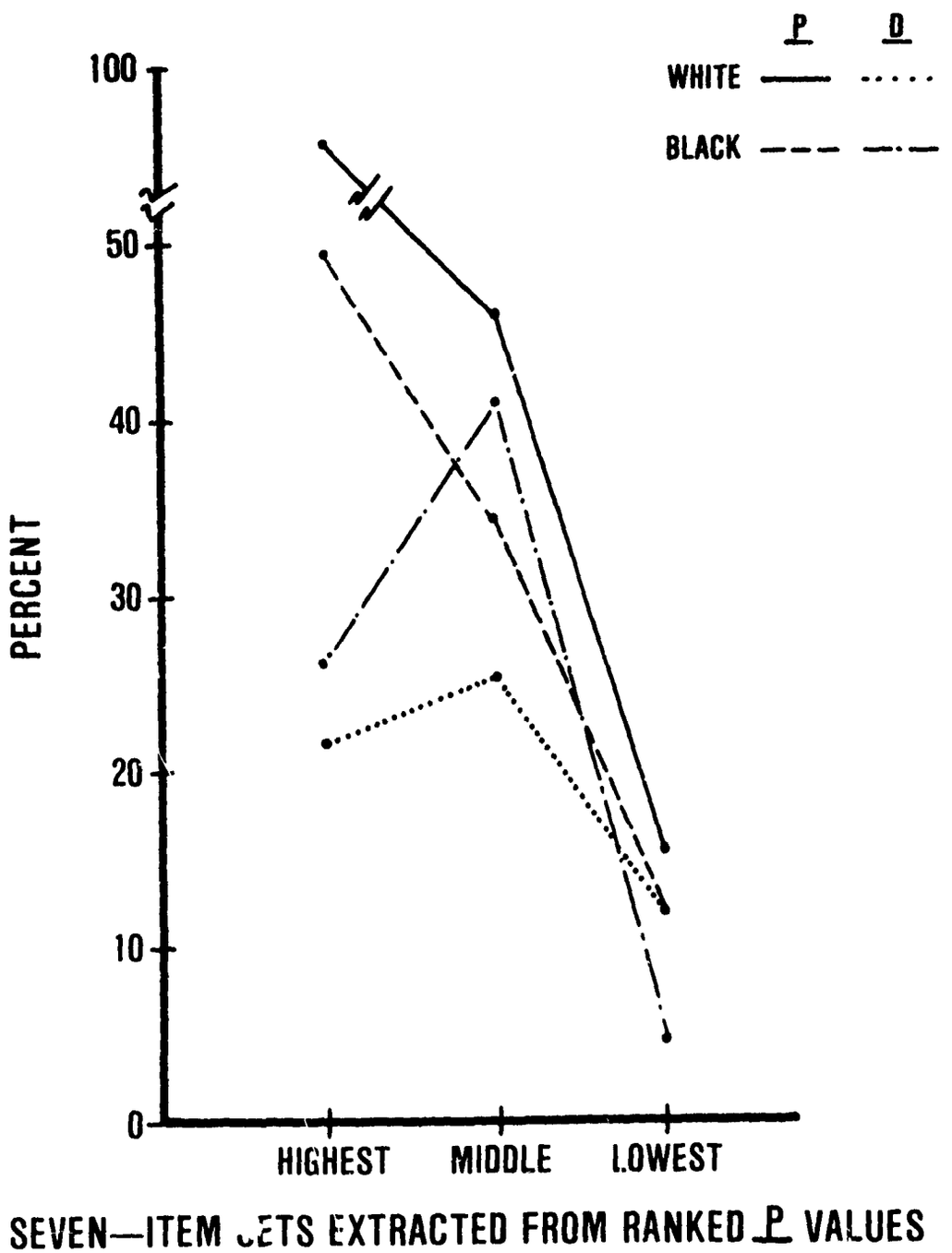


Figure A-1. Median P values and corresponding D values by race (ADJ3 Exam).

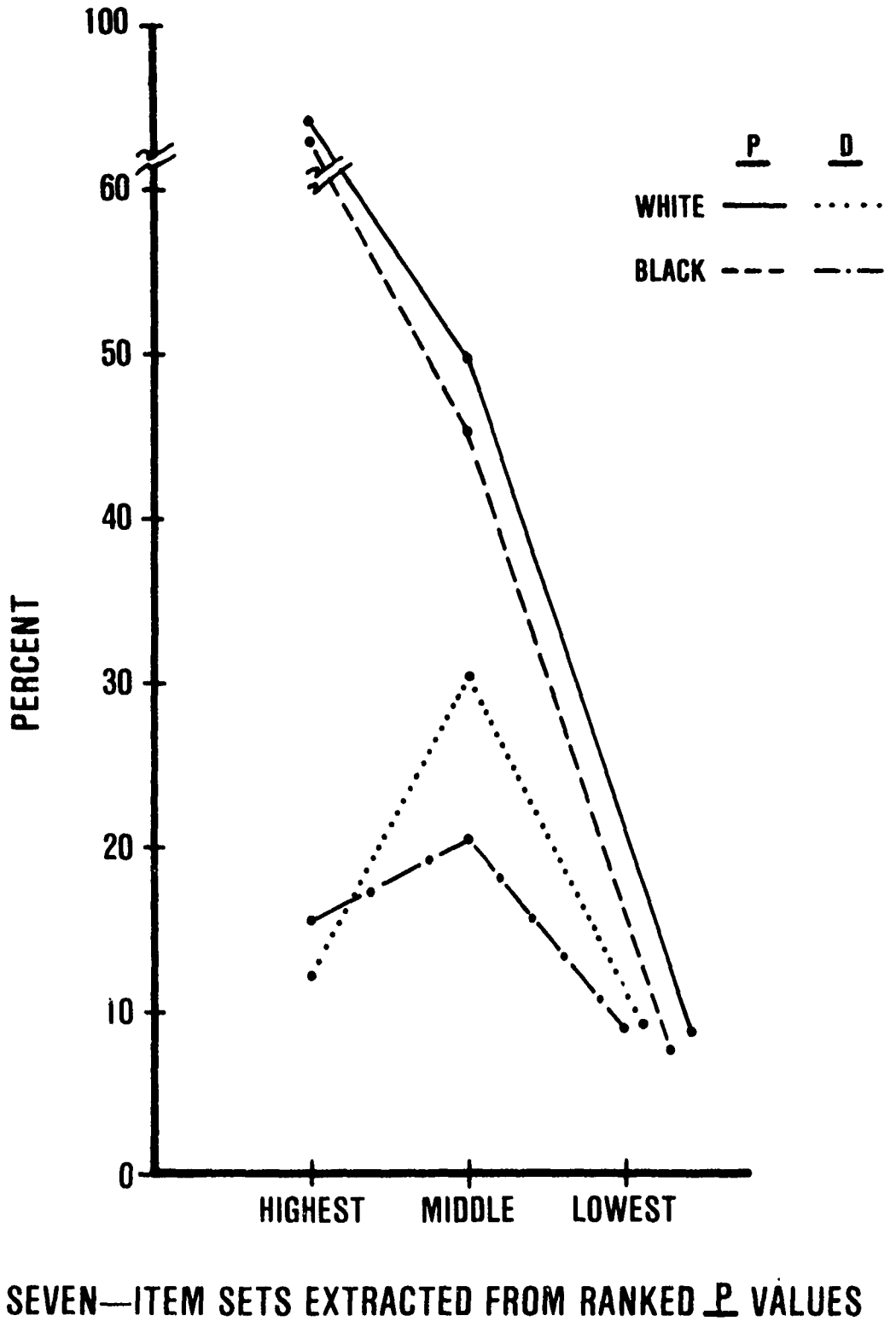


Figure A-2. Median P values and corresponding D values by race (HM3 Exam).

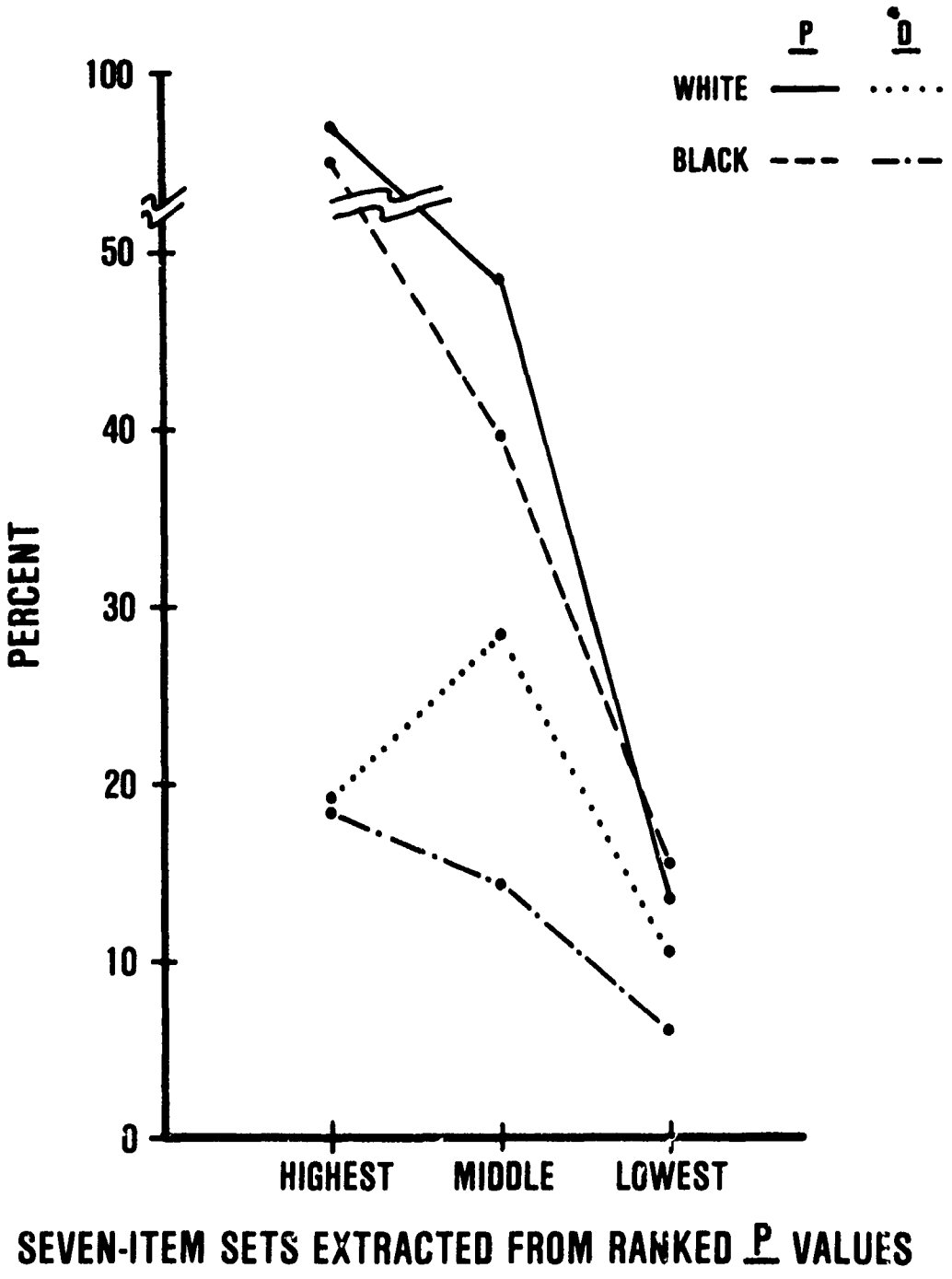


Figure A-3. Median P values and corresponding D values by race (MM3 Exam).

Findings

In the methodological comparisons of alternative measures of item difficulty, the item-score correlations, with the item included in the score were greater than without the item included. With the item in the score, the item-*total* correlation (r_{it}) was greater by about .02 to .04 correlation points, and the item-*section* correlation (r_{is}) was greater by large and varying amounts.

In comparisons of item-section and item-total measures, the percentage difference between high and low scorers answering the item correctly was higher on the item-*section* percentages (D_s values) than on the item-*total* percentages (D_t values) with the item included in both scores.

Item-section (r_{is}) and item-total (r_{it}) correlations, without the item included in the score of either, varied as to which was the larger. Section score intercorrelations within each total test varied from low to high values, suggesting some heterogeneity in some tests or some sections of tests. (Heterogeneity would tend to reduce r_{is} or r_{it} values.) In light of these varying differences, the most useful measure of item differentiation appears to be r_{it} without the item included in the total score.

The P values which corresponded to the highest D values or r_{it} values were higher than the median P values for the total tests, suggesting that easier items might improve item differentiation. In the comparison of the ends of the P value ranges, the highest P values (i.e., easiest items) had corresponding D values which were higher than the corresponding D values of the lowest P values, which suggests that the difficult items are excessively difficult.